

Выборочные функционалы качества классификации

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 10.

Точечные оценки риска

- эмпирический риск,
- скользящий экзамен,
- bootstrap,
- комбинации (несколько статистик),
- другие статистики.

Желательные свойства точечных оценок

- несмещённость,
- состоятельность,
- эффективность.

В отличие от оценивания параметров оценивание риска подразумевает оценивание случайной величины.

Основные понятия

Пусть X – пространство значений переменных,
используемых для прогноза,
 $Y = \{0, 1\}$ – пространство значений прогнозируемых
переменных,
 \mathcal{C} – множество всех вероятностных мер на заданной
 σ -алгебре подмножеств множества $D = X \times Y$.

При каждом $c \in \mathcal{C}$ имеем вероятностное пространство:
 $\langle D, \mathcal{B}, P_c \rangle$, где \mathcal{B} – σ -алгебра, P_c – вероятностная мера.
Параметр c будем называть *стратегией природы*.

Риск

Решающей функцией (алгоритмом классификации) называется соответствие $\lambda: X \rightarrow Y$.

Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}: Y^2 \rightarrow [0, \infty)$.

Положим $\mathcal{L}(y, y') = \begin{cases} 0, & y=y' \\ 1, & y \neq y' \end{cases}$.

Под риском будем понимать средние потери:

$$R(c, \lambda) = \mathbb{E} \mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x)) \, P_c(dx, dy),$$

$x \in X, y \in Y$.

Метод построения решающих функций

Пусть $Q: D^N \rightarrow \Lambda$ — метод (алгоритм) построения решающих функций, $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q , Λ — заданный класс решающих функций.

Метод \tilde{Q} , минимизирующий эмпирический риск, есть

$$\lambda_{\tilde{Q},V} = \arg \min_{\lambda \in \Lambda} \tilde{R}(V, \lambda).$$

Эмпирический риск

Пусть $V = \{(x^i, y^i) \in D \mid i = 1, \dots, N\}$ – случайная независимая выборка из распределения P_c , $V \in D^N$.

Эмпирический риск определим как средние потери на выборке:

$$\tilde{R}(V, \lambda) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y^i, \lambda(x^i)).$$

Функционал, поскольку функция от (решающей) функции (и выборки).

Свойства эмпирического риска

- простота вычисления,
- сильная смещённость,
- состоятельность при соответствующем ограничении на сложность метода,
- малая дисперсия.

Контрольная выборка

Пусть $V^* = \{(x^i, y^i) \in D \mid i = 1, \dots, N^*\}$ – «новая» случайная независимая выборка из распределения P_c , $V^* \in D^{N^*}$.

Оценку риска определим как средние потери на контрольной выборке:

$$R^*(V^*, \lambda) = \frac{1}{N^*} \sum_{i=1}^N \mathcal{L}(y^i, \lambda(x^i)).$$

Доверительный интервал в схеме Бернулли

Односторонний интервал $[0, \hat{p}]$

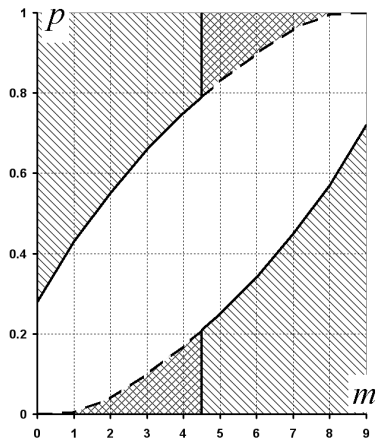
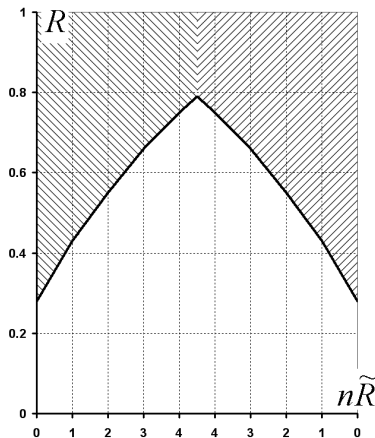
$$\sum_{i=0}^M C_N^i \hat{p}^i \cdot (1 - \hat{p})^{N-i} = \alpha.$$

Двусторонний интервал $[p_1, p_2]$

$$\sum_{i=0}^M C_N^i p_2^i \cdot (1 - p_2)^{N-i} = \sum_{i=M}^N C_N^i p_1^i \cdot (1 - p_1)^{N-i} = \frac{\alpha}{2}.$$

Пример критического множества

Пусть $L = 2$, $\lambda_2(x) = 1 - \lambda_1(x)$, $p = P(y = 0)$, m – количество объектов $y = 0$ в выборке.



Свойства оценки по контрольной выборке

- простота вычисления,
- несмещённость,
- состоятельность,
- известен точный доверительный интервал,
- эффективность (не в том смысле, в котором хотелось бы),
- требуют дополнительной выборки.

Скользящий экзамен

Функционал скользящего экзамена определяется как:

$$\check{R}(V, Q) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y^i, \lambda_{Q, V'_i}(x^i)),$$

где $V'_i = V \setminus \{(x^i, y^i)\}$ — выборка, получаемая из V удалением i -го наблюдения,

Несмещённость скользящего экзамена

Теорема

$$\mathbb{E}\check{R}(V_N, Q) = \mathbb{E}R(V_{N-1}, Q).$$

Доказательство элементарно, хотя неочевидно.

Cross-validaton

K -fold cross-validaton: исходная выборка разбивается на K равных частей (для простоты полагаем, что N кратно K).

$$\check{R}^K(V, Q) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y^i, \lambda_{Q, V_i^K}(x^i)),$$

где V_i^K — выборка, получаемая из V удалением всей подвыборки, которой принадлежит i -е наблюдение.

Разновидности скользящего экзамена

- leave-one-out,
- k -fold crossvalidation,
- случайные подвыборки,
- со стратификацией.

Свойства оценки скользящего экзамена

- относительная простота вычисления,
- несмещённость (не в том смысле, в котором хотелось бы),
- состоятельность (не доказана?),
- большая дисперсия, нет приемлемых оценок доверительного интервала для риска (есть эмпирические свидетельства, что точность сравнима с контролем половинной длины).

Bootstrap

Оценка bootstrap есть

$$\check{R}(V, Q) = \frac{1}{\mathbb{E}|J_0|} \mathbb{E} \sum_{i \in J_0} \mathcal{L}(y^i, \lambda_{Q, \dot{V}}(x^i)),$$

где \dot{V} – выборка, получаемая из V путем N -кратного случайного (равновероятного) выбора ее значений с повторениями, J_0 – множество индексов объектов из V , ни разу не выбранных в \dot{V} , математическое ожидание подразумевает усреднение по выборкам \dot{V} .

Ввиду того, что оценка bootstrap является смещенной, чаще используют ее в комбинации с эмпирическим риском

$$\ddot{R}(V, Q) = e^{-1} \tilde{R}(V, Q) + (1 - e^{-1}) \check{R}(V, Q).$$

Свойства оценки bootstrap

- относительная высокая трудоёмкость вычисления,
- приблизительная несмещённость,
- состоятельность (не доказана?),
- дисперсия неизвестна, но, вероятно, меньше чем у скользящего экзамена.

Гистограммный классификатор

Пусть $X = \{1, \dots, k\}$. Тогда вероятностная мера $P_c[D]$, $c \in C$, задается набором вероятностей

$$\alpha_j = P(x = j), \quad p_j = P(y = 0 \mid x = j).$$

Выборка представляется совокупностью пар

$$V = (v_j \mid j = \overline{1, k}), \quad v_j = (m_j, n_j).$$

Решающая функция минимизирует эмпирический риск независимо в каждой точке $x \in X$: $f(x) = I(m_j < n_j)$.

Выражения для риска

$$\tilde{R}(V) = \sum_{j=1}^k \tilde{r}(m_j, n_j),$$

$$\tilde{r}(m, n) = \frac{1}{N} \tilde{\nu}(m, n), \quad \tilde{\nu}(m, n) = \min(m, n - m);$$

$$R(c, \tilde{\lambda}_{Q,V}) = \sum_{j=1}^k r(m_j, n_j, \alpha_j, p_j),$$

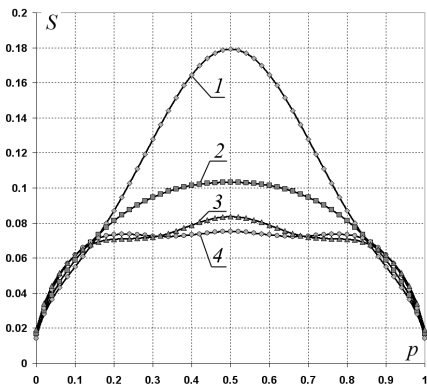
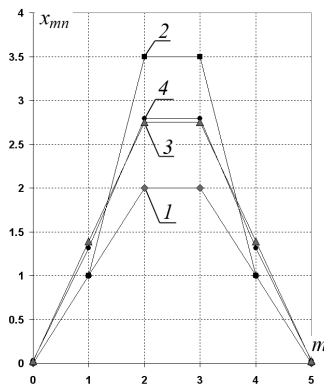
$$r(m, n, \alpha, p) = \alpha \nu(m, n, p),$$

$$\nu(m, n, p) = \begin{cases} 1 - p, & m > n - m; \\ p, & m < n - m; \\ 0,5, & m = n - m. \end{cases}$$

Сравнение оценок

Цифрами обозначены:

- 1 – эмпирический риск, 2 – скользящий экзамен,
3 – комбинированная bootstrap, 4 – оптимизированная.



Качество оценок

В общем случае оценочный функционал — это некоторая функция выборки.

Качество эмпирического функционала $\bar{R}(V, Q)$ как оценки риска обычно характеризуют средним квадратом уклонения, т.е.

$$\Delta = E (\bar{R}(V, Q) - R(c, \lambda_{Q,V}))^2.$$

Существенная проблема заключается в том, что выражения зависят от c — распределения, которое неизвестно.

Кроме того, одно и то же отклонение при разных значениях риска имеет разную значимость.

Доверительный интервал для риска

Доверительный интервал для R зададим в виде $[0, \hat{R}(V)]$, где $\hat{R}(V)$ – оценочная функция или просто оценка (риска). При этом должно выполняться условие:

$$\forall c, P_c(R \leq \hat{R}(V)) \geq \eta,$$

где η – заданная доверительная вероятность.

На практике интервальную оценку будем строить как $\hat{R}(\bar{R}(V))$ – функцию точечной оценки.

Качество интервальной оценки будем характеризовать величиной $E\hat{R}(V)$, которая зависит от c , в виду чего выбор наилучшей оценки становится многокритериальной задачей.

Эмпирические доверительные интервалы для риска

Эмпирический доверительный интервал для R зададим в виде $[0, \hat{R}(\bar{R}(V))]$.

При этом должно выполняться условие:

$$\forall c \in \tilde{C}, P_c(R \leq \hat{R}(V)) \geq \eta,$$

где η – заданная доверительная вероятность, а \tilde{C} – эвристически выбранное множество распределений.

Сравнение интервалов

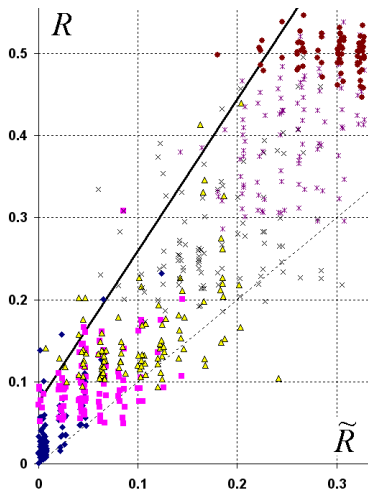
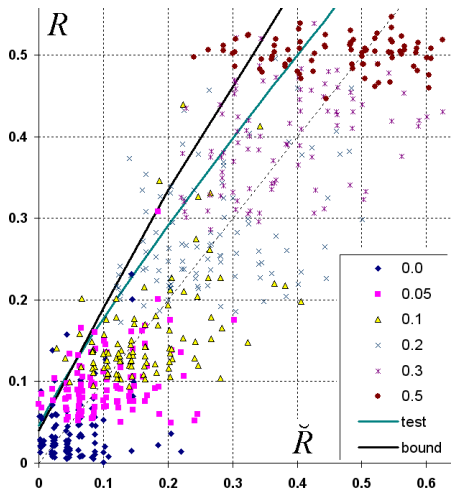
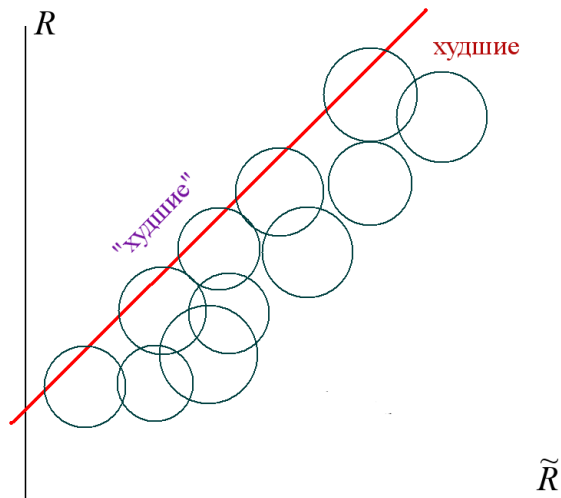


Схема оценивания риска



Замечания

- наилучший способ оценивания риска неизвестен,
- на практике обычно используют скользящий контроль,
- по обучающей выборке нет приемлемых оценок доверительного интервала для риска,
- полезно использовать статистическое моделирование.